Ranking Relevant Verb Phrases Extracted from Historical Text

Eva Pettersson, Beáta Megyesi and Joakim Nivre
Department of Linguistics and Philology
Uppsala University
firstname.lastname@lingfil.uu.se

Abstract

In this paper, we present three approaches to automatic ranking of relevant verb phrases extracted from historical text. These approaches are based on conditional probability, log likelihood ratio, and bag-of-words classification respectively. The aim of the ranking in our study is to present verb phrases that have a high probability of describing work at the top of the results list, but the methods are likely to be applicable to other information needs as well. The results are evaluated by use of three different evaluation metrics: precision at k, R-precision, and average precision. In the best setting, 91 out of the top-100 instances in the list are true positives.

1 Introduction

Automatic analysis of historical text is of great interest not only to the language engineering community, but also to historians and other researchers in humanities, for which historical texts contain information relevant to their research. This information is however not easily accessed. Even in cases where the text has been digitized, contemporary tools for linguistic analysis and information extraction are often not sufficient, since historical text differs in many aspects from modern text, with longer sentences, a different vocabulary, varying word order, and inconsistencies in both spelling and syntax.

In this paper we address the problem of information extraction from historical text, more specifically automatic extraction and ranking of verb phrases describing work. This particular information need has arisen within the *Gender and Work* project, where historians are storing information in a database on what men and women did for a living in the Early Modern Swedish society (i.e. approximately 1550–1800). During this work they have found that working activities in their source material are most often expressed in the form of verb phrases, such as *to fish herring* or *to sell clothes* (Ågren et al., 2011). Our approach to information extraction from historical text, and ranking of the extracted results, is however likely to be applicable to other information needs as well. Furthermore, the methods presented in this paper are not dependent on semantically annotated data, since the only information required is a goldstandard containing positive and negative phrases.

In the ideal case, we would like to extract all verb phrases from a historical text, correctly classify each instance as either describing work or not, and finally present all phrases denoting work, and no other phrases, to the end user. In reality, this is however a tricky task. Even though we have access to a database of phrases previously extracted by the historians as describing work, this does not guarantee that we know how to categorise similar phrases occurring in other texts. For example, the verb *köpa* ('to buy') is sometimes a working activity related to trade, whereas in other contexts, people buy things for non-commercial reasons. In previously unseen texts, there will also most certainly be previously unseen word forms present, which a classifier would not know how to handle. This problem is further aggravated by the high degree of spelling variation in historical text, and inconsistently extracted phrases in the goldstandard (see further Section 3).

Instead of doing a binary classification into phrases denoting work versus phrases not denoting work, we therefore try a ranking approach aiming

Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 39–47,
Beijing, China, July 30, 2015.

to present those verb phrases that most probably describe work at the top of the results list, whereas phrases that are less likely to describe work will be presented further down in the list. In this paper we present three different approaches to verb phrase ranking, based on 1) conditional probability, 2) log likelihood ratio, and 3) bag-of-words classification.

The outline of the paper is as follows. Related work is given in Section 2. In Section 3 we describe the corpus data used in our study. The verb phrase extraction method is presented in Section 4, whereas the ranking methods are described in detail in Section 5. In Section 6, the metrics used for evaluating the ranking approaches are introduced. Finally, the results are presented in Section 7, and conclusions are drawn in Section 8.

2 Related Work

Previous work on information extraction and retrieval from historical text has mainly focused on the problem of searching for certain word forms in historical documents, where spelling variation is challenging.

Baron et al. (2009) addressed the issue of text mining from historical text by developing the VARD 2 tool for automatic translation of historical word forms to a modern spelling as a preprocessing step to text mining. The tool is dictionary-based, and specifically aimed at the Early Modern English language. However, the tool comes with a graphical user interface for interactive semi-automatic adaptation of the tool for handling other language variants as well. They evaluated the adaptability of the tool on Shakespeare's First Folio by first training the tool in the interactive mode on a small sample of the text (5 000 words) corresponding to approximately 6% of the document. Then the proportion of replaced spelling variants was evaluated on the rest of the document, showing an increase from 70.33% for VARD 2 in its original setting to 73.75% after semi-automatic training.

Hauser and Schultz (2007) tried an approach based on weighted edit distance comparisons to match search strings written in Modern German against word forms occurring in documents from the Early New High German period. Pairs of historical word forms and their corresponding modern spelling, retrieved from several lexical sources, were used as training data when learning edit distance weights for commonly occurring differences in spelling between the historical language and the modern language. They showed an increase in information retrieval f-scores for historical tokens from 0.201 without edit distance matching to 0.603 in the best setting.

Pettersson et al. (2013) presented an approach to automatic verb phrase extraction from Early Modern Swedish text. Similar to the methods presented above, the verb phrase extraction process involves a spelling normalisation step, where the historical word forms are translated to a modern spelling, before the extraction of verb phrases is performed. This way, modern taggers and parsers can be used for the linguistic analysis. In their study, the spelling normalisation step is performed by use of character-based statistical machine translation techniques. The verb phrase extraction results showed an increase in the amount of correctly identified verbs from 70.4% for the text in its original spelling to 88.7% in its automatically modernised spelling. Accordingly, the amount of correctly identified complements (including partial matches) increased from 32.9% to 46.2%.

Outside the context of historical data, there is of course a lot of research done on information extraction and data mining, which will not be presented here. Nevertheless, our ranking approaches and the metrics used for evaluating them are inspired by research within this area.

3 Data

In our experiments, we make use of a subset of the Gender and Work (GaW) corpus of Swedish court records and church documents from the Early Modern period. This subset consists of text snippets, referred to as *cases* by the historians. Each case typically contains 4–5 sentences, and comprises at least one phrase describing a working activity. The corpus has been manually analysed by the historians, and those phrases that were judged as denoting work are stored in the GaW database, with information on which case the phrase has been extracted from. This means that we have access both to the source text, and to the phrases within this text that actually describe work. By automatically extracting all the verb phrases from the corpus (see further Section 4 for details on the verb phrase extraction process), it is also possible to infer what verb phrases in the corpus that have **not** been stored in the database, and thus have been

judged **not** do describe work.

This binary classification of verb phrases is of crucial importance to the verb phrase ranking approaches presented in this paper. It is however not a trivial task to decide which of the automatically extracted verb phrases that should be classified as denoting work, when comparing them to the gold-standard of phrases extracted by the historians. Requiring the automatically extracted phrase to be identical to the manually extracted phrase would not be suitable, since the phrases extracted by the historians are not always phrases in the linguistic sense, but may include constituents that would normally be regarded as not belonging to the verb phrase, such as clause adverbials, prepositional phrases, and relative pronouns. Likewise, the manually extracted segments sometimes exclude constituents that would normally be regarded as belonging to the verb phrase, such as indirect objects and adverbial complements. There are also inconsistencies in the spans of the manually extracted phrases, probably partly due to different excerptors.

Similarly, the automatic extraction of verb phrases also results in incomplete verb phrases and phrases containing superfluous constituents. Still, since the overall aim of the verb phrase extraction process is to present elements in the text that may be of interest to the historians, partial phrases and phrases containing extra constituents would still point the user to the right text passage in the source material. Thus, both for training and evaluation we judge an automatically extracted verb phrase as describing work, if there is at least one verb in common between the automatically extracted phrase and the manual excerpt. This means that we run the risk of extracting the wrong instance and still judge it as correct, if there are several instances of the same verb form in the same case. This is especially true for frequent homonyms such as *ha* ('have'), which may be either a temporal auxiliary or a main verb and thus occur several times within the same case or even within the same sentence. In most cases, though, if the automatic excerpt has a verb in common with the manual excerpt, both phrases refer to the same instance. One authentic example from the GaW database is the phrase *sålt een gårdh till hr Leijon Crona* ('sold a farm to Mr Leijon Crona'), which in the automatic excerpt is given as the shorter phrase *sålt een gårdh* ('sold a farm'), but will still be regarded as a true positive (i.e. a phrase describing work).

Even though it has been stated that working activities in the GaW corpus are most often expressed in the form of verb phrases, the phrases stored in the GaW database do not always contain a verb. Common non-verb phrases in the GaW database are noun phrases (*träägårdz dräng på gården*, 'garden servant at the farm'), present participles (*lius säljning*, 'selling of candles'), and past participles or adjectival phrases (*avlönad vid Gripsholm 1572*, 'paid at Gripsholm 1572'). Since our verb-oriented approach explicitly aims at extraction of verb phrases, only phrases in which the tagger is able to identify a verb has been included in our datasets, both for training and for evaluation.

The datasets used in our experiments are presented in Table 3, where *sents* refers to the number of sentences in the corpus, *VPs* are the total amount of verb phrases in the corpus, and *Work VPs* are the amount of these verb phrases that have been judged by the historians as denoting work.

	sents	**VPs**	**Work VPs**
Training	10,623	37,606	10,241
Evaluation	1,358	4,770	1,254

Table 1: Datasets used in our experiments.

As seen from the table, approximately 27% of the verb phrases in the corpus are phrases describing work. It should however be noted that this subset of the corpus is biased towards phrases describing work, since the corpus does not comprise the whole source documents, but only those sections within the documents that actually contain some element describing work.

4 Verb Phrase Extraction

For the task of verb phrase extraction from historical text, we adopt the method introduced by Pettersson et al. (2013), as illustrated in Figure 1.

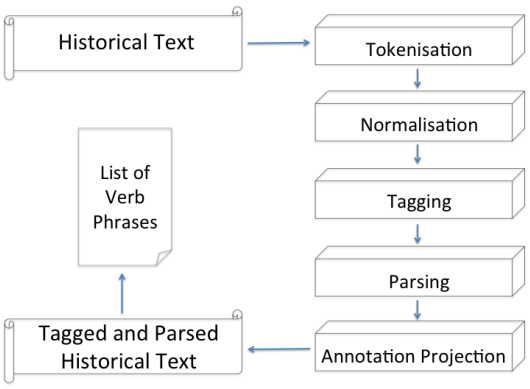

Figure 1: Verb phrase extraction overview.

First, the historical text is tokenised using standard tokenisation methods. The tokenised text is then automatically normalised to a modern spelling, using character-based statistical machine translation methods trained on the same data as described in Pettersson et al. (2013). After spelling normalisation, the modernised text is tagged and parsed using state-of-the-art tools trained for the contemporary language, in this case the HunPOS tagger (Halácsy et al., 2007) with a Swedish model based on the Stockholm-Umeå corpus, SUC (Ejerhed and Källgren, 1997), and the dependency parser MaltParser version 1.7.2 (Nivre et al., 2006a) with a pre-trained model based on the Talbanken section of the Swedish Treebank (Nivre et al., 2006b). Finally, the resulting annotation is projected back to the text in its original, historical spelling. This yields a tagged and parsed version of the historical text in its original spelling, from which the verb phrases are extracted based on the annotation labels.

Using this method, Pettersson et al. (2013) reported an f-score of 88.7% for verb identification, with 46.2% correctly identified complements. In the following, we will focus on the succeeding verb phrase ranking problem, disregarding potential verb phrases that were not found in the extraction process.

5 Verb Phrase Ranking

In the ranking phase, the extracted verb phrases are to be ordered so that those phrases that most probably describe work are presented at the top of the list, and those that most probably do **not** describe work are presented at the bottom of the list. Even though we are focusing on ranking, the training data available is not ranked, but rather classified into phrases describing work and phrases not describing work. This poses special challenges in training the ranking system. We try three different approaches to verb phrase ranking, based on conditional probability, log likelihood calculations, and bag-of-words classification respectively. As a preprocessing step, automatic lemmatisation of the extracted verb phrases (in their automatically modernised spelling) is performed, based on the Saldo dictionary of present-day Swedish word forms (Borin et al., 2008), and the manually lemmatised SUC corpus (Ejerhed and Källgren, 1997).

5.1 Conditional Probability

In the *conditional probability* approach, the probability that a verb phrase describes work, given the verbs present in the phrase, is estimated. For every verb in the phrase to be ranked, the probability that this verb describes a working activity is here estimated using the following formula:

A = number of times the specific verb is part of a verb phrase judged as describing work in the training corpus

B = total frequency of the verb in the training corpus

P(A|B) = $\frac{P(A \cap B)}{B}$

As the final ranking score for the phrase we use either the maximum (i.e. the conditional probability for the verb with the highest conditional probability score), or the average (i.e. the average conditional probability score over all the verbs in the phrase). Furthermore, the conditional probability approach is applied both to purely tokenised data (after spelling normalisation) and to lemmatised data, yielding a total of four different settings for this approach.

5.2 Log Likelihood Ratio

Similar to the conditional probability approach, the *log likelihood* approach also compares the number of times a certain kind of verb phrase has been judged as denoting work to the number of times it has occurred in the corpus without being extracted. One advantage of the log likelihood ratio is however that it also takes into account the number of times a specific token occurs in the corpus, relative to other tokens, rendering a more fair score for low-frequency tokens as compared to high-frequency tokens. We calculate the log likelihood ratio (llr) in accordance with the formula

presented by Dunning (1993), defined as below:

	Event A	**Everything but A**
Event B	k_11: A + B	k_12: B only
Everything but B	k_21: A only	K_22: Neither A nor B

H = Shannon's entropy, computed as the sum of
(k_ij / sum(k)) log (k_ij / sum(k))
llr = 2 sum(k) (H(k) - H(rowSums(k)) - H(colSums(k)))

Applied to the verb phrase ranking problem, the following values are used for the log likelihood variables in order to retrieve a ratio for the likelihood that a certain verb denotes work:

- **k_11**
  The number of times a specific verb occurred in the training corpus and was part of a phrase that the historians extracted as a phrase describing work.
- **k_12**
  The number of times the same verb occurred in the training corpus without being extracted.
- **k_21**
  The number of times any other verb occurred in the training corpus and was part of a phrase that the historians extracted as a phrase describing work.
- **k_22**
  The number of times any other verb occurred in the training corpus without being extracted.

The log likelihood ratio is always given as a positive number. Thus a high number could either mean a high probability that the phrase describes work, or a high probability that the phrase does **not** describe work. For the actual ranking, we have therefore taken into account the relative frequency with which the verb has been judged as describing work in the training corpus, compared to the frequency with which the verb occurred in the training corpus without being extracted. If the verb in question occurs most frequently without being extracted, the log likelihood ratio is prefixed with a minus sign, and treated as representing the probability that the phrase at hand does not describe a working activity. In other cases, the probability score is left as a positive number, thus representing the probability that the phrase at hand actually describes a working activity.

We have tried the following log likelihood settings applied to the ranking problem, where each setting has been tested based on normalised word forms as well as lemmatised data, yielding a total of twelve different settings:

words/lemmas The log likelihood ratio is calculated on the basis of all the tokens (or lemmas) in the phrase. The log likelihood score for the token/lemma with the highest log likelihood ratio is chosen as the ranking score for the whole phrase.

vb The log likelihood ratio is calculated solely on the basis of the verbs in the phrase. The likelihood score for the verb with the highest log likelihood ratio is chosen as the ranking score for the whole phrase.

vbcomp The log likelihood ratio is calculated separately for the verbs and for the non-verb tokens (or lemmas) in the verb phrase. The sum of the maximum verbal log likelihood and the maximum non-verbal log likelihood is chosen as the ranking score for the whole phrase. The hypothesis is that the verbal complements are of importance to distinguish in what contexts a certain verb describes a working activity. For intransitive verbs, only the maximum verbal log likelihood ratio is used for scoring.

vbcomp nn The log likelihood ratio is calculated as in the vbcomp setting, but for the non-verbal calculations, only the nouns are taken into account.

cooc The log likelihood ratio is calculated for the co-occurrence of the verb and each token (or lemma) in the complements. The maximum co-occurrence log likelihood is chosen as the ranking score for the whole phrase. For intransitive verbs, the maximum verbal log likelihood ratio is used for scoring.

cooc nn The log likelihood ratio is calculated as in the cooc setting, but only the nouns in the complements are accounted for.

5.3 Bag-of-Words Classification

In the bag-of-words classification approach, we run a support vector machine (SVM) classifier with the sequential minimal optimization (SMO) algorithm as defined by Platt (1998). All experiments presented here are run with the default linear kernel SVM/SMO settings in the Weka data

mining software package version 3.6.10 (Hall et al., 2009). As training data we use the verb phrases in the training part of the GaW corpus, classified into those that do describe working activities (i.e. have been extracted by the historians) and those that do not describe working activities (consequently those that were not extracted by the historians). We try three different feature selection models for the verb phrase ranking problem, where each model has been applied both to normalised word forms and to lemmatised data, yielding a total of six different settings:

bag of words/lemmas Each word type (or lemma) occurring in the verb phrases in the training corpus is stored as a feature in the model. For every verb phrase to be ranked, each feature is then assigned a value of 1 or 0, depending on weather the specific word form (or lemma) represented by the feature is present in the phrase to be ranked or not.

bag of verbs In the bag-of-verbs setting, only those word forms (or lemmas) that the tagger has analysed as verbs are stored as features. Likewise, only word forms (or lemmas) in the phrase to be ranked that have been analysed as verbs will be compared towards the list of features.

bag of verbs and nouns The bag-of-verbs-and-nouns setting is similar to the bag-of-verbs setting, with the exception that both verbs and nouns are accounted for in this setting. The hypothesis is that the verbal complements, and in particular the nouns occurring in the complements, are of importance to distinguish in what contexts a certain verb describes a working activity.

6 Evaluation

Three different evaluation metrics are applied to the verb phrase ranking results: *precision at k*, *R-precision*, and *average precision*. In accordance with the arguments given in Section 3, an extracted verb phrase is here judged as describing work as long as there is at least one verb in common between the automatically extracted phrase and a manual excerpt from the same case.

6.1 Precision at k

Precision at k is defined as the precision at certain positions in the list of ranked instances (Manning et al., 2008). For example, precision at 10 is the precision achieved for the top-10 instances in the list. For our evaluation, we include precision at 10, 50, and 100 respectively.

6.2 R-precision

R-precision is similar to precision at k, but requires a goldstandard defining the total number of relevant instances. R-precision is then calculated by retrieving the precision score at the position in the list where the number of extracted verb phrases is equal to the number of relevant verb phrases in the goldstandard. At this point, precision and recall are the same, which is why this measure is sometimes also referred to as the *break-even point* (Craswell, 2009). R-precision can be summarised in the following formula:

R = number of relevant phrases in goldstandard
r = extracted relevant phrases at position R
R-precision = $\frac{r}{R}$

In our case we know that the total number of verb phrases denoting work in the evaluation part of the corpus is 1,254. Hence, R-precision is defined as precision at 1,254.

6.3 Average Precision

Average Precision (AVP) is calculated on the basis of the top *n* results in the extracted list, where *n* includes all positions in the list until all relevant instances have been retrieved (Zhang and Zhang, 2009). The average precision can be expressed by the following formula:

r = rank for each relevant instance
P@r = precision at rank r
R = number of relevant phrases in goldstandard
Average precision = $\frac{\sum_r P@r}{R}$

7 Results

7.1 Conditional Probability

Ranking based on conditional probability leads to a substantial improvement in the coverage of verbs denoting work among the top-listed instances, as compared to the baseline case, where the verb phrases are not ranked at all, but simply displayed in the order in which they appear in the source text.

As shown in Table 2, not a single verb phrase describing work is among the top-10 instances

	p10	p50	p100	R-pre	AVP
baseline	0.00	0.10	0.14	0.23	0.24
vb tok avg	0.50	0.66	0.63	0.46	0.44
vb lem avg	0.30	0.64	0.64	0.44	0.43
vb tok max	**0.80**	0.66	0.70	**0.48**	**0.49**
vb lem max	0.60	**0.68**	**0.72**	0.47	0.**49**

Table 2: Results for verb phrase ranking based on conditional probability. p10 = precision at 10, p50 = precision at 50, p100 = precision at 100, R-pre = R-precision, AVP = average precision, baseline = results for the unranked list, tok = token-based model, lem = lemma-based model, avg = probability score based on average value, max = probability score based on maximum value.

without ranking, and only 10% of the top-50 instances are phrases describing work. This could be compared to the token-based model using the maximum value for ranking, where eight out of the top-10 instances are true positives, and 66% of the top-50 instances denote work. At the break-even point (R-precision), nearly half of the positive instances are covered in this setting, as compared to only 23% without ranking. The average precision value follows the R-precision value closely for all settings.

The results also show that ranking based on the highest ranked verb for each phrase, rather than averaging over all the verbs, works the best. Furthermore, we had expected a positive effect of lemmatisation, but interestingly lemmatisation does not help much in the ranking process, and sometimes even lead to lower scores, especially for the models based on average. One reason could be that the kind of documents we are working with (court records and church documents) are almost exclusively written in the past tense, limiting the amount of different verb forms occurring for each lemma. There are also large groups of verbs denoting work, such as *köpa* ('to buy'), *sälja* ('to sell'), *arbeta* ('to work'), *tjäna* ('to serve') etc, that are so commonly occurring in the GaW database that lemmatisation is of little help in the ranking process.

Despite the promising results, there is still room for improvement. The main problem with the conditional probability approach is that no consideration is taken to the number of times a specific verb occurs in the training corpus. Hence, if a certain verb occurs only once in the training corpus, and has been extracted by the historians, it will get the probability 1 of denoting work, and end up at the top of the list. This will be disadvantageous to verbs like *sell* or *buy* that occur many times in the corpus and are often, but not always, extracted by the historians. Likewise, verbs occurring only once without being extracted will always end up at the bottom of the list, together with previously unseen verbs. As discussed in Section 5.2, this skewness is addressed by the log likelihood approach.

7.2 Log Likelihood Ratio

The log likelihood approach, being more sophisticated in balancing the probabilities for low frequency versus high frequency word forms, shows an improvement in the ranking results as compared to the conditional probability approach, as shown in Table 3.

	p10	p50	p100	R-pre	AVP
baseline	0.00	0.10	0.14	0.23	0.24
words	0.80	0.80	0.72	0.52	**0.52**
lemmas	0.60	0.70	0.74	0.45	0.47
vb tok	0.80	0.80	0.72	**0.53**	**0.52**
vb lem	0.50	0.68	0.77	0.51	0.49
vbcomp tok	0.80	**0.84**	**0.83**	0.46	0.49
vbcomp lem	0.80	0.80	0.79	0.45	0.49
vbcomp nn tok	**0.90**	0.82	0.78	**0.53**	**0.52**
vbcomp nn lem	**0.90**	0.82	0.80	0.46	0.49
cooc tok	**0.90**	0.76	0.81	0.36	0.42
cooc lem	0.70	0.74	0.78	0.35	0.40
cooc nn tok	0.50	0.76	0.77	0.31	0.35
cooc nn lem	0.50	0.74	0.77	0.31	0.35

Table 3: Results for verb phrase ranking based on the log likelihood ratio. p10 = precision at 10, p50 = precision at 50, p100 = precision at 100, R-pre = R-precision, AVP = average precision, baseline = results for the unranked list, tok = token-based model, lem = lemma-based model. See Section 5.2 for a description of the other abbreviations used in the table.

It is hard to tell which log likelihood setting is the best, since it depends on what evaluation metric we consider. One option would be to look closer at the results for precision at 100, since it would be a possible scenario to only display the top-100 instances to the user. From these results, we see that the models where the complements are taken into account (*vbcomp* and *cooc* in the table) yield better results than the plain verb-based models. It is also clear that it is more successful to calculate the log likelihood for the verb and the complement separately, and return the sum of these values (*vbcomp*), than to compute a log likelihood score for the co-occurrence of the verb and any of the word forms in the complement (*cooc*).

Furthermore, we get higher precision at k results when we compute the log likelihood for all the word forms in the complement, than when we only consider the nouns in the complement (even though R-precision and average precision

are slightly higher for the noun-restricted settings). A closer look at the top-ranked phrases reveal that they all include the indefinite article, as in *sålt en* ('sold a'), *köpt en* ('bought a'), *skjutit en* ('shot a'), *stulit en* ('stolen a'), etc. This is logical in a way, since it indicates that it is of greater importance to the log likelihood ratio that **something** is sold or bought or worked with etc, than exactly **what** is sold or bought or worked with, where the latter would be better expressed by the nouns in the complement than by the indefinite article.

7.3 Bag-of-Words Classification

The ranking results for the bag-of-words classification approach are presented in Table 4.

	p10	p50	p100	R-pre	AVP
baseline	0.00	0.10	0.14	0.23	0.24
words	0.60	0.88	0.84	0.49	0.53
lemmas	0.50	0.82	0.81	0.49	0.52
vb tok	**1.00**	0.92	0.87	**0.52**	**0.55**
vb lem	**1.00**	**0.94**	0.85	0.50	0.53
vbnn tok	0.80	0.92	**0.91**	0.50	0.54
vbnn lem	0.70	0.92	0.88	0.50	0.54

Table 4: Results for verb phrase ranking based on machine learning. p10 = precision at 10, p50 = precision at 50, p100 = precision at 100, R-pre = R-precision, AVP = average precision, baseline = results for the unranked list, words = bag of words, lemmas = bag of lemmas, vb = bag of verbs, vbnn = bag of verbs and nouns, tok = token-based model, lem = lemma-based model.

The results are generally higher than for both the conditional probability method and the log likelihood calculations. For the best precision at 100 results, 91% of the instances are verb phrases describing work. Similar to the results for conditional probability and log likelihood ratio, lemmatisation generally has no positive effect on the results. Unlike the results for the log likelihood approach though, it seems beneficial to exclude non-nouns from the complements in the machine learning approach. This is however only true for the precision at 100 metric, whereas the other metrics indicate the opposite.

7.4 Summary of the Results

Table 5 summarises the results for the methods with the highest precision at 100 score within the three different approaches. As seen from the table, the bag-of-words classification approach yields the highest score for every evaluation metric used when comparing these results.

	p10	p50	p100	R-pre	AVP
baseline	0.00	0.10	0.14	0.23	0.24
cond prob	0.60	0.68	0.72	0.47	0.49
llr	**0.80**	0.84	0.83	0.46	0.49
bow	**0.80**	**0.92**	**0.91**	**0.50**	**0.54**

Table 5: Summary of the results for verb phrase ranking. p10 = precision at 10, p50 = precision at 50, p100 = precision at 100, R-pre = R-precision, AVP = average precision, baseline = results for the unranked list, cond prob = conditional probability, llr = log likelihood, bow = bag-of-words classification.

8 Conclusion

In this paper we have presented three approaches to ranking of relevant verb phrases extracted from historical text, based on 1) conditional probability, 2) log likelihood ratio, and 3) bag-of-words classification. Neither of the methods are dependent on semantically annotated data, since they all rely on binary classified training data of verb phrases containing the desired information versus other verb phrases.

Even though the ranking systems were trained on binary data rather than ranked data, all three methods yield very promising results. The bag-of-words classification approach reaches the highest scores according to all three evaluation metrics used (precision at k, R-precision, and average precision). The best bag-of-words setting is token-based (as opposed to lemma-based), taking both the verbs and the nouns in the verb phrases into account in the ranking process. In this setting, 91% of the top-100 instances in the results list are true positives.

Although the experiments were conducted for the specific task of extracting and ranking verb phrases describing work in historical Swedish text, the methods developed are language-independent and could easily be applied to other languages and information needs by simply altering the training data. It would therefore be interesting to evaluate the presented ranking methods on other information needs, document types, source languages, and time periods etc. Future work also includes a user-based evaluation together with the historians. The outcome of such an evaluation would not only show to what degree the system is useful in the extraction process, but also whether the phrases stored in the database will be different in any way when using our tool for extraction as compared to a fully manual extraction process, for instance regarding consistency.

References

Maria Ågren, Rosemarie Fiebranz, Erik Lindberg, and Jonas Lindström. 2011. Making verbs count. The research project 'Gender and Work' and its methodology. *Scandinavian Economic History Review*, 59(3):273–293.

Alistair Baron, Paul Rayson, and Dan Archer. 2009. Automatic standardization of spelling for historical text mining. In *Proceedings of Digital Humanities*.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. Saldo 1.0 (svenskt associationslexikon version 2). Språkbanken, University of Gothenburg.

Nick Craswell. 2009. R-precision. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 2453–2453. Springer US.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Eva Ejerhed and Gunnel Källgren. 1997. Stockholm Umeå Corpus. Version 1.0. Produced by Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University. ISBN 91-7191-348-3.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 209–212, Prague, Czech Republic.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explorations*, 11:1.

Andreas Hauser and Klaus Schultz. 2007. Unsupervised learning of edit distance weights for retrieving historical spelling variations. In *Proceedings of FSTAS 2007*, pages 1–7.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006a. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC)*, pages 2216–2219, Genoa, Italy, May.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006b. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC)*, pages 24–26, Genoa, Italy, May.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2013. An SMT approach to automatic annotation of historical text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA. NEALT Proceedings Series 18; LinkÃűping Electronic Conference Proceedings.*, volume 87, pages 54–69.

John C. Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Advances in Kernel Methods - Support Vector Learning.

Ethan Zhang and Yi Zhang. 2009. Average precision. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 192–193. Springer US.